子音の音韻類似度を用いた 併置型駄洒落検出手法の有効性について

谷津 元樹　　荒木 健治
北海道大学 情報科学研究科
{motoki.yatsu,araki}@ist.hokudai.ac.jp

1 はじめに

近年，機械がユーモアを表出・理解することに対する関心が高まっている．ユーモアへの関心は，機械によるユーモアがそれに接するユーザの生活の質 (QOL) を向上させることに対する関心と人工知能技術全体の発展への貢献に対するものに大きく分けられる．

ユーモアを使用する人の QOL 向上を実証する例として，社会心理学においては，ユーモアを交えることによるアサーティブネス (自己表現・意見表明) の向上 [1]，臨床心理学においては，日常生活におけるユーモアの表出によるメンタルヘルスの改善への寄与を示す報告 [2] がある．

ユーモアの理解を行うには，最初にユーモアそのものを高い精度で検出する必要がある．言語的なユーモアには，散文の形で記述されたユーモアと韻文のユーモアがある[1]．

散文ユーモアの検出に関しては Twitter[2]等より取得した口語文及び散文ユーモアのコーパスを対象に研究が行われている [3] が，韻文のユーモアに対しては，規模の大きなコーパスに基づく検出手法に関する研究は，著者らの知るところでは確認されていない．

韻文ユーモアとして代表的な**駄洒落**においては，一部の音素が異なるが音韻的に類似性の高い 2 つの区間 (**種表現**および**変形表現**) を持つもの (**併置型駄洒落**) が多くを占めていることが，**2.2** で述べる調査によって判明している．併置型駄洒落の検出を行うには，種表現および変形表現の正確かつ音の脱落などを許容する検出手法を用いる必要がある．

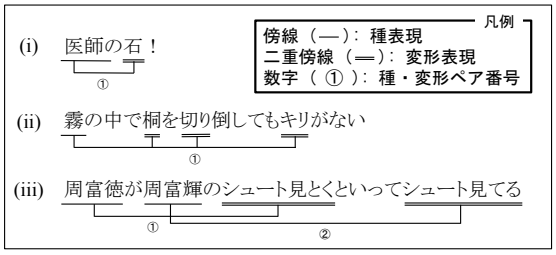

図 1: 併置型駄洒落における種表現と変形表現の例

次に本稿の構成を述べる．**1** 章では駄洒落検出の必要性について述べた．続いて **2** 章では，検出対象となる併置型駄洒落の類型及び基本構造，そして検出において要求される条件について述べる．**3** 章で本稿での分析に用いたデータについて述べる．**4** 章では提案する音素ペアの類似性の算出手法について述べる．**5** 章では提案手法を用いた駄洒落内の種表現・変形表現の検出実験について述べ，結果に対する考察を行う．

2 併置型駄洒落の検出

本章では，併置型駄洒落を含む，駄洒落の類型について述べ，併置型駄洒落の検出に必要となる併置型駄洒落の構造に関する知識について述べる．

2.1 駄洒落の主な類型

駄洒落は，音韻的な類似や一致関係により対となる 2 つの音素列を含む表現と捉えられる．滝澤 [4] の行った形態的な定義では，併置型駄洒落および**重畳型駄洒落**の 2 種に分類される．前者においては類似する対となる音素列が対象文内に共起するが，後者ではしないという差異がある．共起する場合，入力が単文であっても駄洒落を含むかについての識別が可能となる．種表現および変形表現の現れ方については **2.3** で述べる．

[1] 散文ユーモアは，例えば以下のような笑い話となる．

> 我が家は一軒家だが、実は貸家である。先日、屋根修理の事で大家が来たとき、つい「汚い家ですがどうぞ」と言ってしまった。(出典：http://matome.naver.jp/odai/2129478948138388901)

また，韻文のユーモアは，本稿で取り上げる駄洒落のほか，謎掛けの多くがこれにあたる．

[2] https://twitter.com

表 1: 駄洒落の類型とその構成比

併置型	563 (全体 93.8%)
その他の類型	33 (全体 5.5%)
取得失敗	4 (全体 0.7%)
Imperfect な併置型駄洒落	521 (全体 86.8%, 併置型 92.5%)

Kawahara[5] はこの併置型駄洒落において，種表現と変形表現の音韻が完全に一致するものを **perfect** な駄洒落，語音の一部の変化または脱落により一致しないものを **imperfect** なものと呼称している．

(iv) 相談してもいいそうだん!

(v) ハンペンを食べる時は、シャンペン!

上記の (iv) は perfect な駄洒落，(v) は imperfect な駄洒落の例である．本研究においては，語音の変化に柔軟な検出器を実装することを目的に，perfect/imperfect 双方の駄洒落を検出の対象とする．

2.2 各類型の構成比

2.1 で述べた 2 種の類型のそれぞれが，駄洒落の標本中に占める割合を調査した．調査者は第一著者である．標本は，**3** 章で述べる分析対象として取得した全駄洒落文より，600 項目を無作為抽出した．ここで得られた駄洒落の類型別の構成を表 1 に示す．

この結果から，併置型駄洒落が全体において高い割合 (93.8%) を占めること，中でも一部音素が異なるが音韻的に類似した 2 区間を持つ imperfect な駄洒落の比率が駄洒落文全体の 86.8%と大きいことがわかる．

また，imperfect な駄洒落が検出対象とする併置型駄洒落の 90%超を占めることから，音韻の類似部分の検出においては語音の異なりを吸収することが必要と考えられる．

2.3 併置型駄洒落の構造

併置型駄洒落は文内に 2 つの音韻的に類似した区間 (種表現および変形表現) をもつ．併置型駄洒落において，種表現は文内の独立形態素 (名詞・動詞・形容詞・感動詞で自立語となるもの) あるいはその句 (名詞句・動詞句・形容詞句) を，変形表現は種表現との音韻類似度を持った文内の任意の区間を指す．

1 つの併置型駄洒落において，種表現および変形表現の対が少なくとも 1 対出現する．図 1 の文 (i) および (ii) はその例である．ただし，文 (ii) のように，1 つの種表現に対し複数の変形表現が対応する場合がある．

種・変形表現がいずれも交換可能な場合，便宜上先行する表現を種表現とし，その他をこれに対応する変形表現とする．

2.4 モーラ単位音素列による表現

本稿において処理の対象とする文は，すべて読み仮名をモーラ単位に分割し，それぞれの子音音素と母音音素を分離後配列化する．本来単独では音価[3]をもたない拗音 (「ゃ」,「ゅ」,「ょ」および「ゎ」) 文字が検出の際に直前の読み仮名と切り離されて一致することを避けるためである．長音は直前の母音の反復として扱い，また促音 (「っ」) および撥音 (「ん」) モーラに独自の記号を与える．例えば「インターショップ」に対しては {/*,i/,/*,N/,/t,a/,/*,a/,/sh,o/,/*,Q/,/p,u/} という配列に変換される．ここで，N は撥音，Q は促音を指す．その他の子音部の音素の表記はヘボン式ローマ字による音声表記と同一である．

3 分析に用いたデータ

日常会話における駄洒落の使用に関しては，実コーパスやその統計的分析結果がこれまでに得られておらず，具体的な実用例においての検出を行うことが困難なため，正例データの収集を Web より行った．結果，6,675 文の分析用のデータを取得した．データの取得源である Web サイトを表 2 に示す．

3.1 データの前処理

3.3.1 読み仮名列の抽出

形態素解析器 MeCab[4] ならびに補助形態素解析器として形態素解析器 JUMAN[5] を用いて読み仮名列を抽出する．補助形態素解析器として JUMAN を用いた理由は，MeCab が読みを抽出できなかった単語に対し，異なる形態素解析手法および辞書を用いて読みの抽出を再度試みるためである．

3.3.2 モーラ音素列への変換

駄洒落を含む各文を **2.4** に示した形式に変換する．その際，**3.3.1** で抽出した読み仮名列を入力として用いる．

4 音素ペアの音韻類似度の算出

本章では，2 つの表現における同一位置の音節における子音または母音同士の音素ペアのもつ音韻的な類似性を，音韻類似度として数量化する手法について述べる．

[3] 表記に対応する音素．

[4] http://chasen.org/〜taku/software/mecab

[5] http://nlp.ist.i.kyoto-u.ac.jp/index.php?JUMAN

表 2: 駄洒落を含む文の取得元の Web サイト

サイト名	URL	取得件数
ダジャレナビ	http://www.dajarenavi.net	4,927
Dajare Station	http://dajare.jp	1,107
ダジャレネット	http://www.dajare.net	204
ひとくちダジャレ大集合	http://www.biwa.ne.jp/~aki-ina/gyagu.html	134
ダジャレ集 ダジャレ事典	http://dajareshuu.web.fc2.com	123
ダジャレの缶詰	http://www.geocities.jp/pikumin_hiroba/dajare.html	73
駄洒落倶楽部	http://with2.net/dajakura	54
ダジャレ広場	http://www1.ocn.ne.jp/~origo/dazyare	39
駄洒落を言ったのは誰じゃ?	http://wtpage.info/dajare	14

4.1 子音および母音の類似性

Kawahara らの研究 [5] においては，子音の類似性スコアを，2 つの表現における同一位置の音節の母音が一致する場合に生起した子音音素ペアの対数 O/E 比を計算することにより得ている．具体的には，**4.2** に述べる手法によりスコアを算出する．

Kawahara ら [6] はまた，母音音素ペアの類似性スコアを，同様に対数 O/E 比を求めることにより算出している．したがって，母音の類似性を子音の類似性と複合的に用いることが可能と考えられる．しかしながら，子音の音韻類似度のみを用いる場合に比べアライメントの手法が複雑化するため，実装は今後の課題とする．

4.2 音韻類似度の計算

子音同士の音韻類似度の指標として，対数 O/E 比を用いる．対数 O/E 比は，Kawahara らによる子音の心理音響的類似性の指摘 [7] の他，広く音韻論の分野で 2 音素の共起の度合の指標として用いられる．対数 O/E 比は，式 (1) により定義される．

$$OER(p_1, p_2) = \log \frac{N_{p_1,p_2} + N_{p_2,p_1}}{N_{\text{pair}} \cdot P(p_1) \cdot P(p_2)} + c \quad (1)$$

ただし， p, q は異なる子音を，N_{p_1,p_2} は p_1, p_2 の順に表れる子音の対 $[p_1, p_2]$ の対象コーパスにおける生起頻度を，c は正の実定数を指す．計算において p_1, p_2 は交換可能である．

音韻類似度は，種表現および変形表現の境界情報の付与されたアノテーションデータより事前に算出が可能である．この場合，アノテーションにおいて開始位置が揃えられている種表現・変形表現において位置の一致する音素同士を音素ペアとする．

4.3 音韻類似度を用いた変形表現検出

駄洒落を含む文を検出するには，種表現に対し処理対象の文全体より，類似部分となる変形表現を検出する必要がある．具体的には，次のいずれかの条件を満たすモーラ音素列区間を変形表現とする．

(A) 種表現との間の音素ペアの類似度の総和が閾値を上回る．

(B) 種表現との間で音素列が完全に一致している．

(A) は具体的には，種表現候補および変形表現候補における子音音素ペアの音韻類似度の総和が閾値を上回る場合，その変形表現候補を変形表現として検出するものである．音韻類似度は，開発データに存在するすべての子音音素ペアについて、式 (1) を用いて事前に算出した対数 O/E 比を用いる．**(B)** は最も単純な類似区間の検出手法であり，性能比較用の手法として **5.3** にて使用する．

(A) を適用した例として，文 (v) における種表現となる形態素「人魚」と変形表現となる区間「人情」の間の音韻類似度を計算する．

(v) 人魚の人情劇！

{/n,i/, /*,N/, /gy,o/, /n,o/, /n,i/, /*,N/, /j,o/, /*,u/, /g,e/, /k,i/}

子音間類似度を異なる子音に対して計算する．ここでは子音のペア [gy-j] の対数 O/E 比を利用する．**5.1** で構築したアノテーションデータを用いた場合，スコアの総和は 0.2 となる．

5 変形表現の検出性能の比較

4 章において述べた子音の音韻類似度を適用した駄洒落を含む文の検出実験を行い，音韻類似度の有効性を検証する．

表 3: 各条件による駄洒落検出実験の結果

条件	精度	再現率	F 値
(A)	0.651	0.712	0.680
(B)	0.645	0.699	0.671

5.1 種・変形表現アノテーション

分析用データの駄洒落を含む 6,675 文より 675 文を無作為抽出し，抽出した 675 文に対し **3.1** で述べたデータの前処理を施したものについて，種表現および変形表現のモーラ単位音素列上の境界の付与を人手により行った．次に，残りの 6,000 文を評価用データとして分離した．

5.2 負例データの準備

駄洒落を含む文の検出を行うために，駄洒落を含まない文のデータセットを構築した．データは 2011 年より 2012 年の間に Ameba ブログ[6]において書かれたブログ記事を収集した YACIS コーパス [8] より 6,000 件を，文長の制約[7]とともに抽出した．

結果として，駄洒落を含む文 (正例) 6,000 文および，同数の駄洒落を含まない文 (負例) を準備した．

5.3 検出実験および結果

5.2 において準備した正例・負例文に対して，**4.3** において述べた，変形表現となる音素列区間の検出条件 **(A)** および **(B)** を用いた駄洒落の検出実験を行った．対数 O/E 比の算出は **5.1** にて構築したアノテーションデータを用いた．実験の結果を表 3 に示す．

ここで，条件 **(A)** における閾値は，**5.1** のアノテーションデータ 675 文を用いて同様の実験を行った結果が最良である 0.06 とした．また，各性能値の計算方法を式 (3)〜(5) に示す．

$$\text{精度} = \frac{\text{検出成功した駄洒落を含む文の数}}{\text{駄洒落を含む文とした判定の総数}} \quad (3)$$

$$\text{再現率} = \frac{\text{検出成功した駄洒落を含む文の数}}{\text{駄洒落を含む文の総数}} \quad (4)$$

$$F\text{ 値} = \frac{2 \times \text{精度} \times \text{再現率}}{\text{精度} + \text{再現率}} \quad (5)$$

[6] http://ameblo.jp

[7] 駄洒落文の文長が正規分布に従うと仮定し，駄洒落音素列長の平均 μ_l および標準偏差 σ_l に対し，式 (2) の範囲内の文長 l をもつ文のみを抽出した．

$$\mu_l - 2\sigma_l \leq l \leq \mu_l + 2\sigma_l \quad (2)$$

5.4 考察

音韻類似度を用いた手法 **(A)** の性能は，完全一致手法 **(B)** に対して精度 0.006 ポイント，再現率 0.012 ポイント，F 値においては 0.009 ポイントの向上が確認された．これらの結果より，音素ペアの音韻類似度を考慮することによってより正確な駄洒落の検出が可能となることが確認された．

6 おわりに

本稿では，併置型駄洒落の検出に向け，内容語形態素の種表現に対し音韻的に一致する，あるいは類似性の高い区間を検出する手法の提案を行った．コーパスより算出した音素間類似度を用いた種・変形表現の検出実験の結果，子音の音素ペア同士の音韻類似度を考慮することにより検出性能が向上することから，子音の音韻類似度を考慮することの有効性が確認された．

今後の課題としては，**4.2** で指摘した母音を用いた音素ペア音韻類似度の導入に加え、音素間類似性スコアを算出する際に用いる種・変形表現境界の付与済みコーパスのさらなる拡充等が挙げられる．

参考文献

[1] 山田浩平, 朝野聡, 物部博文. 対人葛藤場面での断り行動に対する自己効力感と社会的スキル及びアサーティブな態度，ユーモア対処との関わり. 学校保健研究, Vol. 54, No. 3, pp. 203–210, 2012.

[2] 塚脇涼太, 深田博己, 樋口匡貴. ユーモア表出が表出者自身の不安および抑うつに及ぼす影響過程. 実験社会心理学研究, Vol. 51, No. 1, pp. 43–51, 2011.

[3] 天谷祐介, ラファウジェプカ, 荒木健治. 単語間類似度を用いた物語ユーモア認識手法の性能評価. 人工知能学会第 2 種研究会 ことば工学研究会資料, pp. 63–69, 2013.

[4] 滝澤修. 記述された「併置型駄洒落」の音素上の性質. 自然言語処理, Vol. 2, No. 2, pp. 3–22, 1995.

[5] S. Kawahara. Probing knowledge of similarity through puns. *Proceedings of Sophia University Linguistic Society*, Vol. 23, pp. 110–137, 2009.

[6] S. Kawahara and K. Shinohara. Calculating vocalic similarity through puns. *Journal of the Phonetic Society of Japan*, Vol. 13, pp. 101–110, 2009.

[7] S. Kawahara and K. Shinohara. The role of psychoacoustic similarity in Japanese puns: A corpus study. *Journal of Linguistics*, Vol. 45, pp. 111–138, 2009.

[8] M. Ptaszynski, P. Dybala, R. Rzepka, K. Araki, and Y. Momouchi. YACIS: A Five-Billion-Word Corpus of Japanese Blogs Fully Annotated with Syntactic and Affective Information. In *Proceedings of The AISB/IACAP – LaCATODA workshop*, pp. 40–49, 2012.